

ABSTRACT OF THE DISCLOSURE

A document information processing apparatus includes a form element analyzer (12) for performing a form element analysis on a plain document inputted from a plain document input unit (10) by using a dictionary stored in a dictionary storage unit so as to decompose the plain document into tokens, a syntax analyzer (13) for analyzing the part of speech of each of the tokens obtained by the form element analyzer so as to generate a structured document containing meaningful words, an element refinement processing unit (15) for performing a markup process of adding data associated with each of the meaningful words included in the structured document generated by the syntax analyzer and stored in a data storage unit (14) to each of the meaningful words so as to generate a markup document, and a markup document output unit (17) for outputting the markup document generated by the element refinement processing unit.